REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*
25-7-2008

2. REPORT TYPE
Final Technical Report

3. DATES COVERED *(From - To)*
1-4-2007-31-3-2008

4. TITLE AND SUBTITLE
(U) (DURIP 07) Terascale Cluster for Advanced Turbulent Combustion Simulations

5a. CONTRACT NUMBER

5b. GRANT NUMBER
FA9550-07-01-0288

5c. PROGRAM ELEMENT NUMBER
61103F

6. AUTHOR(S)
Stephen B. Pope and Steven R. Lantz

5d. PROJECT NUMBER
5094

5e. TASK NUMBER
US

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Cornell University
Upson Hall
Ithaca NY 14853

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AFOSR/NA
875 Randolph St
Suite 325, Room 3112
Arlington VA 22203-1768

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

A terascale computing system was purchased and brought into full operation. The main cluster consisted of 36 dual-dual-core Dell servers connected by a QLogic InfiniBand switch. The full system included 2 interactive servers holding 2 terabytes of file storage. The cluster enabled combustion simulations to run on up to 140 processor cores. On benchmarking tests with HP Linpack, the system was found to perform well, achieving 1.15 teraflop/s, or 68.7% of the theoretical peak based on the cluster's 3.0-GHz Intel processors. Initial tests on codes used in turbulent combustion research showed excellent scalability. The system served as an excellent resource for turbulent combustion research and for the training of graduate students in high-performance computing.

15. SUBJECT TERMS
Turbulent combustion, cluster

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
					Julian M. Tishkoff
a. REPORT	b. ABSTRACT	c. THIS PAGE	UL	8	19b. TELEPHONE NUMBER *(include area code)*
Unclassified	Unclassified	Unclassified			(703) 696 8478

Final Performance Report

Terascale cluster for advanced turbulent combustion simulations

Defense University Research Instrumentation Program (DURIP)

Grant No.: FA9550-07-01-0288

Period of award: 4/1/07 – 3/31/08

PI: Stephen B. Pope
Mechanical & Aeorpace Engineering
Cornell University

Co-PI: Steven R. Lantz
Center for Advanced Computing
Cornell University

July 25, 2008

Abstract

A terascale computing system was purchased and brought into full operation. The main cluster consisted of 36 dual-dual-core Dell servers connected by a QLogic InfiniBand switch. The full system included 2 interactive servers holding 2 terabytes of file storage. The cluster enabled combustion simulations to run on up to 140 processor cores. On benchmarking tests with HP Linpack, the system was found to perform well, achieving 1.15 teraflop/s, or 68.7% of the theoretical peak based on the cluster's 3.0-GHz Intel processors. Initial tests on codes used in turbulent combustion research showed excellent scalability. The system served as an excellent resource for turbulent combustion research and for the training of graduate students in high-performance computing.

1. Introduction

In May 2007 we received an award of $171,315 under the Defense University Research Instrumentation Program (DURIP) for the purchase of a terascale computing cluster to advance the research on turbulent combustion in the PI's research group. The purchase has been completed, and the cluster is a success in all respects.

In the next section we describe the hardware purchased and the use of the funds. Then the performance of the cluster is described both for benchmarking tests and for initial research applications. The system achieves 1.15 teraflop/s (68.7% of the theoretical peak), and in turbulent combustion applications it displays excellent parallel scaling.

The cluster is heavily utilized by the graduate students in the PI's research group, including an NDSEG fellow, and Ph.D. students supported by the PI's AFOSR grant.

The cluster is housed and maintained by the Cornell Center for Advanced Computing. We have received a gift of $10,000 from Microsoft Corporation towards the cost of maintenance.

2. Terascale cluster

2.1 Description of the system

We have given the name CATS (for Combustion And Turbulence Simulator) to the terascale system that was obtained through this grant. CATS is a 36-node Dell cluster featuring two dual-core Intel Xeon 5160 (3GHz "Woodcrest") processors per node, tied together using a QLogic 4X SDR InfiniBand interconnect. CATS includes an interactive login node and a file server, each holding in excess of 1 terabyte of file storage. The 35 active compute nodes of CATS enable us to run up to 140-core parallel MPI batch jobs; one node is reserved to run the scheduler. CATS is operated and maintained for our group by the Cornell Center for Advanced Computing (CAC).

As its operating system, CATS runs Microsoft's Compute Cluster System (CCS), which consists of Windows Server 2003 plus the Compute Cluster Pack. The latter includes MS-MPI (a derivative of MPICH2) and the Job Manager. Microsoft has agreed to make a donation of software and additional support to help with a planned upgrade to HPC Server 2008. One interactive node is designated as the login node, and it has a full suite of analysis and development tools, including Microsoft Visual Studio, Intel Fortran, and Matlab. Fluent from Ansys, Inc. is installed on the login node and all compute nodes.

2.2 Use of the funds

The following table details expenditures made under this grant.

Bill of Materials			
Product Description	**Qty**	**Unit Cost**	**Cost**
DELL Hardware			
Nodes (Compute and Interactive)			
PowerEdge 1950, 2x Intel Xeon 5160 (3 GHz), 8 GB RAM - set of 4	9	12,888.00	115,992.00
PowerEdge 2950, 2x Intel Xeon 5160 (3 GHz), 16 GB RAM - set of 2*	1	11,516.24	11,516.24
(*each includes 5x 300 GB HDD configured as a RAID-5 array)			
Racks & Rack Components			
Dell Rack 4210 - pair	1	1,737.92	1,737.92
Dell PowerEdge 2161 Digital KVM	1	3,328.13	3,328.13
44 1U close out panels (to occupy empty slots for proper airflow)	1	879.56	879.56
1U KMM (keyboard, mouse, monitor)	1	599.00	599.00
Subtotal			6,544.61
Interconnect			
High Speed/Compute Fabric			
QLogic 4X SDR InfiniBand Interconnect	1	32,903.00	32,903.00
Admin Fabric			
Dell PowerConnect 6248, 48 Port 1GbE Managed Switch	1	1,158.70	1,158.70
Hardware Total			**168,114.55**
Installation Costs			
Machine room installation by CAC staff			3,200.45
Total Materials and Installation			**171,315.00**

The system as delivered was significantly enhanced compared to the one we originally proposed. This was due to market-driven price reductions in many of the components, as well as negotiations that took place with Dell following award notification in April 2007.

Of particular note is a 20% reduction in the cost of the IB interconnect. This was made possible by the use of a modular 36-port QLogic switch in place of the (underpopulated) 48-port Cisco switch in the proposal. Note too that in the QLogic solution, the number of InfiniBand HCAs and cables increased from 32 to 36. In the end, only 20% of the hardware cost was devoted to the high-performance interconnect for the cluster, as compared to 25% in the initial proposal, without sacrificing performance.

The above cost savings enabled the expansion of the cluster from 32 compute nodes to 36, along with the addition of two other servers: an interactive development (login) node plus a dedicated file server, each holding over 1 TB of RAID-5 file storage.

Under Dell's standard arrangement with CAC, all equipment is covered by three-year warranties that are valid through 2010. Extended warranties on the InfiniBand equipment were paid for by other means, as required by the terms of the contract.

3. Performance of the cluster

3.1 Benchmarking

CATS has been tested with several standard benchmarks to assure proper functioning of the hardware and software. Point-to-point bandwidth and latency tests were conducted by the co-PI using the OSU Micro-Benchmark (OMB) suite. Following the installation of the latest OpenIB drivers for the QLogic HCAs, the results come very close to the hardware limits: 950 MB/s bandwidth and 8 microseconds latency. These measurements compare very favorably with the ideal results, which would be 1 GB/s bandwidth and a few microseconds latency for 4X SDR InfiniBand.

Capability testing of the full cluster was conducted using the High-Performance Linpack (HPL) benchmark. Careful tuning of input parameters is typically needed to generate optimal results with this benchmark. Invaluable assistance from Microsoft during the tuning process is gratefully acknowledged. The best result achieved was 1.15 Tflop/s out of an ideal 1.68 Tflop/s, where the latter assumes that all 140 cores schedule and execute 4 flops every cycle. The 1.15 Tflop/s result is nearly 70% of peak, which is acceptable performance for an InfiniBand cluster of this size.

3.2 Scaling studies

CATS is presently being used to aid the development of a new parallel code for simulations of turbulent combustion. The development plan includes an explicit goal of scaling up to large numbers of processor cores. Fluid turbulence in the code is modeled via the Large Eddy Simulation (LES) approach, and in order to make fast progress, software from Stanford University has been adapted for development purposes. The Stanford code was already parallelized with MPI, so some preliminary testing has been done with it to see how well it scales on CATS. Here, “scaling” is used in the strong sense of parallel speedup for a given problem size; the ideal is represented by a straight line of slope 1.

Results from the initial testing of this LES code are shown in Figure 1. The test case is a calculation of DLR Flame A using the flamelet model. The pressure equation is solved using a multigrid technique. A regular mesh of 128x128x64 is imposed in the x-r-theta cylindrical coordinate space. Domain decomposition is done primarily in the x direction, with secondary decomposition in the r direction. The calculation is restarted from a statistically stationary state, and the solution is advanced for another 1000 time steps.

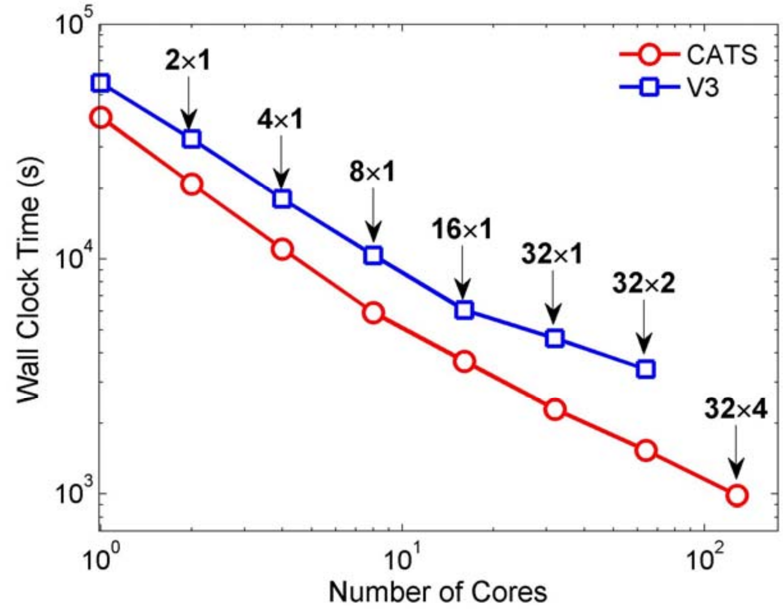

Figure 1: Wall clock time of the LES code on CATS and V3

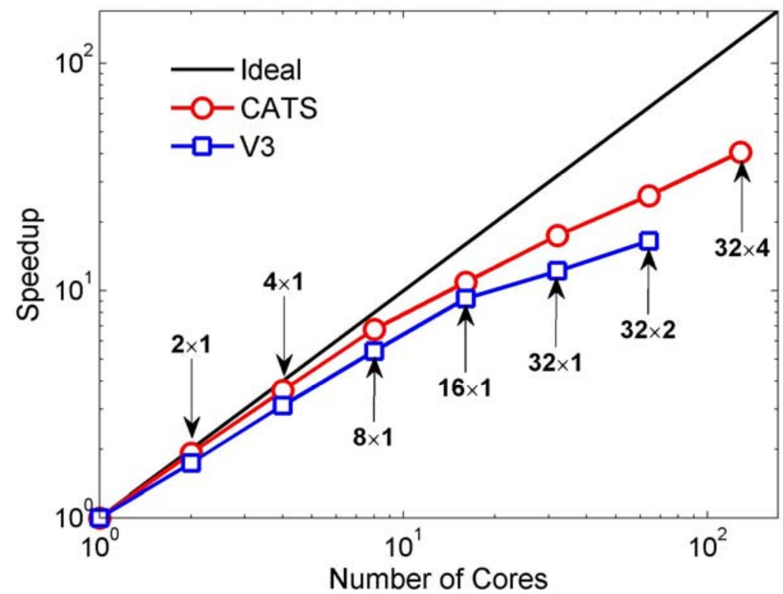

Figure 2: Parallel speedup of the LES code on CATS and V3

The figure clearly shows that the per-core performance of CATS is superior to that of V3, a predecessor cluster at CAC with a gigabit Ethernet interconnect (note, both of these clusters feature fully non-blocking switches). At each core count, identical domain decompositions were used for the test; these are denoted in the figure as MxN, where M is the number of x-partitions, and N is the number of r-partitions. In every case, the solution is achieved in less time on CATS. This is a surprisingly good result given that the clock speed of the Intel Xeon processors on V3 is 3.6 GHz, compared to 3.0 GHz on CATS. The better performance on CATS can be attributed to other characteristics of its newer Xeon processors: higher memory bandwidth, larger cache sizes, and the ability to execute more flops/cycle.

In Figure 2, the data from Figure 1 have been re-plotted for each cluster to show the parallel speedup, which is defined as the ratio of the single-core time to the multi-core time. The figure shows that as expected, the parallel efficiency of the LES code begins to tail off as the core count approaches 100. Such scaling behavior is not unusual; in the context of the overall code development project, this scaling is acceptable because the LES portion of the computation is not projected to be the dominant cost. In any event, for present purposes, the key point of Figure 2 is that CATS shows better parallel scaling than V3. Any other outcome would have been surprising since InfiniBand ought to outperform gigabit Ethernet in every respect. Taken together, Figures 1 and 2 demonstrate that CATS is an excellent resource for simulating fluid turbulence using a distributed-memory MPI-parallel code.

Combustion chemistry in the emerging code is simulated using the Filtered Density Functional or FDF methodology. To accelerate computations in the FDF portion of the code, the main algorithm utilized by the research group is ISAT, for In Situ Adaptive Tabulation, a technique which was developed by the PI and his collaborators. ISAT has been parallelized using our x2f_mpi software library, which was designed with ISAT in mind.

Preliminary scaling studies of parallel ISAT were conducted for the case of a partially-stirred reactor (PaSR). The results, exhibited in Figure 2 below, indicate good scalability on the CATS cluster. Here, “scalability” is used in the more typical sense of maintaining a constant time-to-solution when the problem size is increased in proportion to the core count. Even when message passing is maximized by re-distributing particles uniformly and randomly (URAN) at every step, the performance is comparable to that of purely local processing (PLP), where there is no message passing. This indicates that many particles can be quickly passed between processes while others are being evaluated in ISAT. The message-passing cost must be accordingly low. Thus, the measurements give good evidence that the IB interconnect in CATS is effective in holding down parallel overhead as both the core count and the total number of particles increase.

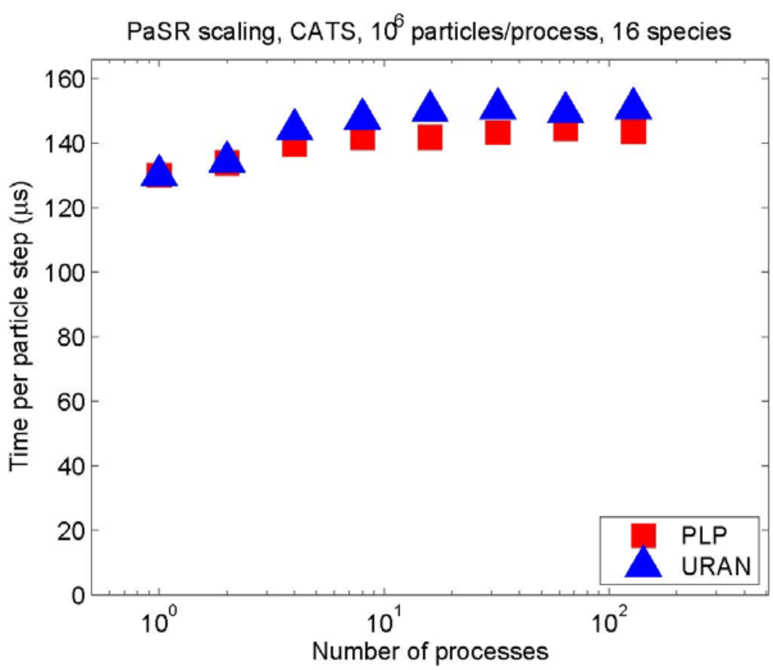

Figure 2: Constancy of PaSR particle-step time vs. core count on CATS

4. Conclusion

The program has been completed successfully and all funds expended. The cluster performs extremely well, including in tests of our turbulent combustion codes. The cluster provides an extremely valuable resource to accelerate our research progress over the next 3-5 years.